Uncertainty in Statistical Inference – Part 2

Frank Schorfheide
University of Pennsylvania

November 4, 2024

Uncertainty in Statistical Inference

- Much of what you have seen in econometrics classes is about attaching standard errors (s.e.) to estimates.
- They can be used to construct coverage intervals, e.g.,

$$\text{point estimate} \pm 2 \times s.e. \tag{1}$$

- The s.e. are supposed to summarize uncertainty associated with your estimates.
- **Our question:** Do they? In what sense?

Grouping Empirical Studies

- $N \cdot (J+1)$ studies that try to estimate a similar parameter.
- Use ij subscripts.
- $i = 1, \ldots, N$ groups of studies.
- Baseline studies (for group i): $j = 1, \ldots, J$.
- Validation study (for group i): $j = J+1$.
- Notation similar to panel forecasting problem: previously i was bank, t was time period, T periods to for estimation, forecast outcome in $T+1$.

Linking Parameters Across Studies

- **Bayesian** hierarchical modeling assumption

$$\theta_{ij}|(\tau_i, \nu_i) \overset{iid}{\sim} \mathcal{N}(\tau_i, \nu). \tag{2}$$

- Parameters are highly correlated but not identical.
- Parameters in each group share common mean τ_i.
- If $\nu = 0$ parameters in each group are identical.

Parameter Estimates and S.E.

- Study ij reports point estimate $\hat{\theta}^{ij}$ and s.e. σ_{ij}.
- Give estimate a frequentist interpretation to obtain quasi likelihood:

$$\hat{\theta}_{ij}|\theta_{ij} \sim \mathcal{N}\left(\theta_{ij}, \sigma_{ij}^2\right). \tag{3}$$

- **Likelihood function** for estimates:

$$\hat{\theta}_{ij}|(\tau_i, \nu) \overset{iid}{\sim} \mathcal{N}\left(\tau_i, \nu + \sigma_{ij}^2\right). \tag{4}$$

$$p\left(\hat{\theta}_{i,1:J}|\tau_i, \nu\right) \propto \prod_{j=1}^{J}(\nu + \sigma_{ij}^2)^{-1/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\frac{(\hat{\theta}_{ij} - \tau_i)^2}{\nu + \sigma_{ij}^2}\right\}, \tag{5}$$

Posterior Distribution of τ_i

- **Bayes Theorem:**

$$p(\tau_i|\hat{\theta}_{i,1:J},\nu) \propto p\big(\hat{\theta}_{i,1:J}|\tau_i,\nu\big)p(\tau_i). \tag{6}$$

- Improper (does not integrate to one) prior $p(\tau_i) \propto c$.

- Solving the square in the exponential term of (5), we obtain

$$\sum_{j=1}^{J}\frac{(\hat{\theta}_{ij}-\tau_i)^2}{\nu_i+\sigma_{ij}^2} = \sum_{j=1}^{J}\frac{\hat{\theta}_{ij}^2}{\nu_i+\sigma_{ij}^2} - 2\tau_i\sum_{j=1}^{J}\frac{\hat{\theta}_{ij}}{\nu_i+\sigma_{ij}^2} + \tau_i^2\sum_{j=1}^{J}\frac{1}{\nu_i+\sigma_{ij}^2}.$$

- The posterior takes the form

$$\tau_i|\big(\hat{\theta}_{i,1:J},\nu_i\big) \sim \mathcal{N}\big(\bar{\tau}_i,\bar{V}_{\tau_i}\big), \quad \bar{V}_{\tau_i} = \frac{1}{J}\left(\frac{1}{J}\sum_{j=1}^{J}\frac{1}{\nu_i+\sigma_{ij}^2}\right)^{-1}, \quad \bar{\tau}_i = \bar{V}_{\tau_i}\left(\sum_{j=1}^{J}\frac{\hat{\theta}_{ij}}{\nu_i+\sigma_{ij}^2}\right). \tag{7}$$

Special Cases:

1. Single baseline study ($J = 1$):

$$\bar{V}_{\tau_i} = \nu + \sigma_{ij}^2, \quad \bar{\tau}_i = \sigma_{ij}^2. \tag{8}$$

2. Identical parameters across studies ($\nu = 0$):

$$\bar{V}_{\tau_i} = \frac{1}{J}\left(\frac{1}{J}\sum_{j=1}^{J}\frac{1}{\sigma_{ij}^2}\right)^{-1}, \quad \bar{\tau}_i = \bar{V}_{\tau_i}\left(\sum_{j=1}^{J}\frac{\hat{\theta}_{ij}}{\sigma_{ij}^2}\right). \tag{9}$$

Predictive Distribution

- **Goal:** create credible interval for $\hat{\theta}_{i,J+1}$ from baseline studies; assess its coverage.
- Assume (similar to baseline studies):

$$\hat{\theta}_{i,J+1}|(\tau_i, \nu_i) \sim \mathcal{N}\big(\tau_i, \nu_i + \omega_i^2\big). \tag{10}$$

- Define

$$\bar{V}_{\hat{\theta}_{i,J+1}} = \bar{V}_{\tau_i} + \nu + \omega_i. \tag{11}$$

- The $1-\alpha$ predictive interval is

$$CI^{\hat{\theta}_{i,J+1}}\big(\{\hat{\theta}_{i,1:J}\}, \nu_i\big) = \left[\bar{\tau}_i - z_{\alpha/2}\sqrt{\bar{V}_{\hat{\theta}_{i,J+1}}},\ \bar{\tau}_i + z_{\alpha/2}\sqrt{\bar{V}_{\hat{\theta}_{i,J+1}}}\right]. \tag{12}$$

- Just as in interval forecast evaluation, compute empirical coverage frequency:

$$\text{CovFreq}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big\{\hat{\theta}_{i,J+1} \in CI^{\hat{\theta}_{i,J+1}}\big(\{\hat{\theta}_{(i,\cdot)}\},\nu_i\big)\Big\}. \tag{13}$$

- Rewrite the expression for the coverage frequency:

$$\begin{aligned}\text{CovFreq}_N \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big\{ & -z_{\alpha/2}\sqrt{\bar{V}_{\hat{\theta}_{i,J+1}}} \leq (\hat{\theta}_{i,J+1}-\theta_{i,J+1}) \\ & + (\theta_{i,J+1}-\tau_i) + (\tau_i - \bar{\tau}_i) \leq z_{\alpha/2}\sqrt{\bar{V}_{\hat{\theta}_{i,J+1}}}\Big\}.\end{aligned}$$

- Frequentist calculation: condition on $\{\tau_i\}_{i=1}^N$ and average over $(\hat{\theta}_{i,J+1} - \theta_{i,J+1})$ and $\hat{\theta}_{i,1:J}$:

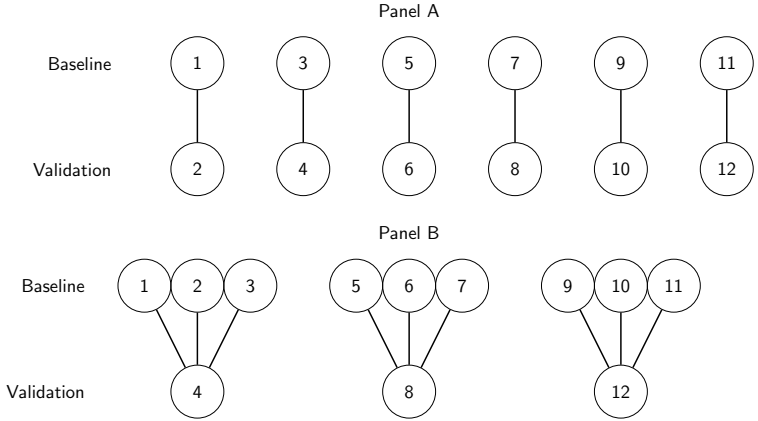
$$\begin{aligned} \bar{\tau}_i - \tau_i &= \sum_{j=1}^{J} \frac{\bar{V}_{\tau_i}}{\nu_i + \sigma_{ij}^2} \hat{\theta}_{ij} - \sum_{j=1}^{J} \frac{\bar{V}_{\tau_i}}{\nu_i + \sigma_{ij}^2} \tau_i \\ &= \sum_{j=1}^{J} \frac{\bar{V}_{\tau_i}}{\nu_i + \sigma_{ij}^2} (\hat{\theta}_{ij} - \theta_{i,j} + \theta_{i,j} - \tau_i) \\ &\sim \mathcal{N}\left(0,\ \bar{V}_{\tau_i}\right). \end{aligned} \tag{14}$$

  Here we used that $\sum_{j=1}^{J} \bar{V}_{\tau_i}/(\nu + \sigma_{ij}^2) = 1$ and $\mathbb{V}\left[\hat{\theta}_{ij} - \theta_{i,j} + \theta_{i,j} - \tau_i\right] = \sigma_{ij}^2 + \nu$.

- Deduce that the empirical coverage frequency converges to the nominal coverage probability as $N \longrightarrow \infty$.

Odds and Ends

- Estimation of hyperparameter ν, then replace ν by $\hat{\nu}$ (empirical Bayes).
- Effect of using a standard error estimate instead of true s.e.
- Fixed number of studies available, choose N and J.



Monte Carlo Simulation

Param.	Description
N_{sim}	Number of Monte Carlo repetitions
N	Number of validation studies
J	Number of baseline studies for each validation study
λ	Variance of τ_i
ν	Variance of θ_{ij} conditional on τ_i
σ^2	Sampling variance of estimator $\hat{\theta}_{ij}$ given θ_{ij}
π_κ	Fraction of studies with distorted standard errors
φ	Distortion factor for standard error

Potentially Distorted S.E.

- Set $\sigma_{ij}^2 = \omega^2 = \sigma^2$ and $\nu_i = \nu$.
- π_κ and φ control the distortion of the reported standard errors.
- We assume that

$$\kappa_{ij} \overset{iid}{\sim} \text{Bernoulli}(\pi). \tag{15}$$

- Estimator is distributed as

$$\hat{\theta}_{ij} | \theta_{ij} \sim \mathcal{N}(\theta_{ij}, \sigma^2). \tag{16}$$

- Econometrician reports

$$\hat{\sigma}_{ij}^2 = \sigma^2 \Big(1 + \kappa_{ij}(\varphi - 1) \Big). \tag{17}$$

Data Generation

The following calculations are repeated N_{sim} times:
Generate parameter estimates and standard error estimates for the $N(J+1)$ studies: for $i = 1, \ldots, N$

1. Generate $\tau_i \overset{iid}{\sim} \mathcal{N}(0, \lambda)$.
2. Draw $\hat{\theta}_{ij}$, $j = 1, \ldots, J$, from (4), and $\hat{\theta}_{i,J+1}$ from (10), i.e. from a $\mathcal{N}(\tau_i, \nu + \sigma^2)$ under the homoskedastic design.
3. Draw κ_{ij} and the distorted squared standard errors $\hat{\sigma}^2_{ij}$, $j = 1, \ldots, J+1$, according to (15) and (17).

Simulation: Use True ν

	N=50		$N = 500$		$N = 5,000$	
	CovFreq	p-value	CovFreq	p-value	CovFreq	p-value
			$\pi_k = 0$, $\varphi = 0$			
$J = 2$	0.80	0.56	0.80	0.54	0.80	0.53
$J = 4$	0.80	0.55	0.80	0.55	0.80	0.50
$J = 9$	0.80	0.55	0.80	0.54	0.80	0.49
			$\pi_k = 0.5$, $\varphi = 0.75^2$			
$J = 2$	0.76	0.47	0.77	0.25	0.77	0.00
$J = 4$	0.77	0.50	0.77	0.23	0.77	0.00
$J = 9$	0.77	0.49	0.77	0.21	0.77	0.00
			$\pi_k = 0.5$, $\varphi = 1.25^2$			
$J = 2$	0.82	0.55	0.83	0.25	0.83	0.00
$J = 4$	0.83	0.54	0.83	0.25	0.83	0.00
$J = 9$	0.83	0.55	0.83	0.25	0.83	0.00

Notes: The results are based on $N_{sim} = 500$ Monte Carlo Repetitions. The coverage frequency is computed across the N_{sim} repetitions. The nominal coverage of the intervals is 80%. The p-value is computed for the null hypothesis that that the coverage probability equals its nominal value.

Simulation: Use Estimated ν

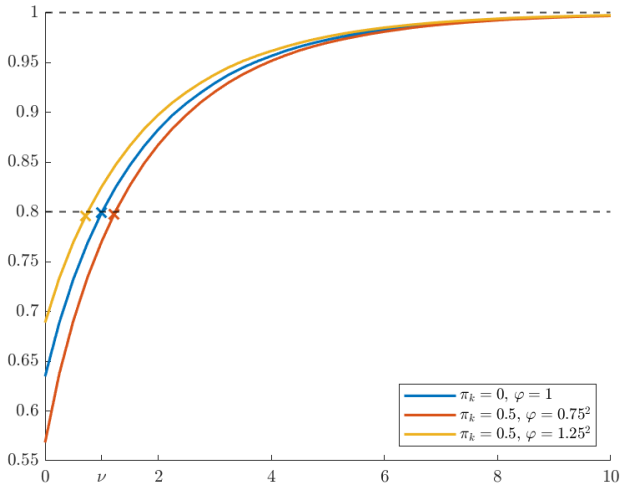
	N=50			$N = 500$			$N = 5,000$		
	CovFreq	p-value	$\hat{\nu}$ Bias	CovFreq	p-value	$\hat{\nu}$ Bias	CovFreq	p-value	$\hat{\nu}$ Bias
				$\pi_k = 0$, $\varphi = 0$					
$J = 2$	0.79	0.47	-0.03	0.80	0.47	0.00	0.80	0.44	0.00
$J = 4$	0.80	0.52	0.01	0.80	0.51	0.00	0.80	0.46	0.00
$J = 9$	0.80	0.55	0.01	0.80	0.52	0.00	0.80	0.48	0.00
				$\pi_k = 0.5$, $\varphi = 0.75^2$					
$J = 2$	0.78	0.48	0.19	0.80	0.47	0.22	0.80	0.45	0.22
$J = 4$	0.80	0.52	0.23	0.80	0.50	0.22	0.80	0.45	0.22
$J = 9$	0.80	0.54	0.23	0.80	0.53	0.22	0.80	0.47	0.22
				$\pi_k = 0.5$, $\varphi = 1.25^2$					
$J = 2$	0.78	0.47	-0.31	0.80	0.46	-0.28	0.80	0.42	-0.28
$J = 4$	0.80	0.53	-0.27	0.80	0.51	-0.28	0.80	0.42	-0.28
$J = 9$	0.80	0.55	-0.27	0.80	0.51	-0.28	0.80	0.46	-0.28

Notes: The results are based on $N_{sim} = 500$ Monte Carlo Repetitions. The coverage frequency and the bias of $\hat{\nu}$ are computed across the N_{sim} Monte Carlo repetitions. The nominal coverage of the intervals is 80%. The p-value is computed for the null hypothesis that that the coverage probability equals its nominal value.

What Happened Here?

- With true ν: we are able to detect that some intervals are incorrectly specified.
- With estimate ν: estimate adjusts so that coverage frequency matches nominal level.
- Oooops...
- Think of ν as controlling the degree of external validity, i.e., heterogeneity in θ_{ij} across j.
- If intervals are too small, we can reduce the degree of external validity to get the right coverage.
- **Thus, we can only assess the coverage statements, conditional on assumptions on the degree of external validity.**

Simulation: Coverage as a Function of ν



Let's Try This on Actual Estimates: Many Labs Replication

Replication

Investigating Variation in Replicability

A "Many Labs" Replication Project

Richard A. Klein,[1] Kate A. Ratliff,[1] Michelangelo Vianello,[2] Reginald B. Adams Jr.,[3] Štěpán Bahník,[4] Michael J. Bernstein,[5] Konrad Bocian,[6] Mark J. Brandt,[7] Beach Brooks,[1] Claudia Chloe Brumbaugh,[8] Zeynep Cemalcilar,[9] Jesse Chandler,[10,36] Winnee Cheong,[11] William E. Davis,[12] Thierry Devos,[13] Matthew Eisner,[10] Natalia Frankowska,[6] David Furrow,[15] Elisa Maria Galliani,[2] Fred Hasselman,[16,17] Joshua A. Hicks,[12] James F. Hovermale,[1] S. Jane Hunt,[18] Jeffrey R. Huntsinger,[19] Hans IJzerman,[7] Melissa-Sue John,[20] Jennifer A. Joy-Gaba,[17] Heather Barry Kappes,[21] Lacy E. Krueger,[18] Jaime Kurtz,[22] Carmel A. Levitan,[23] Robyn K. Mallett,[19] Wendy L. Morris,[24] Anthony J. Nelson,[3] Jason A. Nier,[25] Grant Packard,[26] Ronaldo Pilati,[27] Abraham M. Rutchick,[28] Kathleen Schmidt,[29] Jeanine L. Skorinko,[22] Robert Smith,[14] Troy G. Steiner,[4] Justin Storbeck,[8] Lyn M. Van Swol,[30] Donna Thompson,[15] A. E. van 't Veer,[7] Leigh Ann Vaughn,[31] Marek Vranka,[32] Aaron L. Wichman,[33] Julie A. Woodzicka,[34] and Brian A. Nosek[29,35]

[1]University of Florida, Gainesville, FL, USA. [2]University of Padua, Italy. [3]The Pennsylvania State University, University Park, PA, USA. [4]University of Würzburg, Germany. [5]Pennsylvania State University Abington, PA, USA. [6]University of Social Sciences and Humanities Campus Sopot, Poland. [7]Tilburg University, The Netherlands. [8]City University of New York, USA. [9]Koç University, Istanbul, Turkey. [10]University of Michigan, Ann Arbor, MI, USA. [11]HELP University, Kuala Lumpur, Malaysia. [12]Texas A&M University, College Station, TX, USA. [13]San Diego State University, CA, USA. [14]Ohio State University, Columbus, OH, USA. [15]Mount Saint Vincent University, Nova Scotia, Canada. [16]Radboud University Nijmegen, The Netherlands. [17]Virginia Commonwealth University, Richmond, VA, USA. [18]Texas A&M University-Commerce, TX, USA. [19]Loyola University Chicago, IL, USA. [20]Worcester Polytechnic Institute, MA, USA. [21]London School of Economics and Political Science, London, UK. [22]James Madison University, Harrisonburg, VA, USA. [23]Occidental College, Los Angeles, CA, USA. [24]McDaniel College, Westminster, MD, USA. [25]Connecticut College, New London, CT, USA. [26]Wilfrid Laurier University, Waterloo, ON, Canada. [27]University of Brasília, DF, Brazil. [28]California State University, Northridge, CA, USA. [29]University of Virginia, Charlottesville, VA, USA. [30]University of Wisconsin-Madison, WI, USA. [31]Ithaca College, NY, USA. [32]Charles University, Prague, Czech Republic. [33]Western Kentucky University, Bowling Green, KY, USA. [34]Washington and Lee University, Lexington, VA, USA. [35]Center for Open Science, Charlottesville, VA, USA. [36]PRIME Research, Ann Arbor, MI, USA. [37]University Nijmegen, The Netherlands

Abstract. Although replication is a central tenet of science, direct replications are rare in psychology. This research tested variation in the replicability of 13 classic and contemporary effects across 36 independent samples totaling 6,344 participants. In the aggregate, 10 effects replicated consistently. One effect – imagined contact reducing prejudice – showed weak support for replicability. And two effects – flag priming influencing conservatism and currency priming influencing system justification – did not replicate. We compared whether the conditions such as lab versus online or US versus international sample predicted effect magnitudes. By and large they did not. The results of this small sample of effects suggest that replicability is more dependent on the effect itself than on the sample and setting used to investigate the effect.

Keywords: replication, reproducibility, generalizability, cross-cultural, variation

Replication is a central tenet of science; its purpose is to confirm the accuracy of empirical findings, clarify the conditions under which an effect can be observed, and estimate the true effect size (Brandt et al., 2013; Open Science Collaboration, 2012, 2014). Successful replication of an experiment requires the recreation of the essential conditions of the initial experiment. This is often easier said than done. There may be an enormous number of variables

Social Psychology 2014; Vol. 45(3):142–152
DOI: 10.1027/1864-9335/a000178

Let's Try This on Actual Estimates: Many Labs Replication

Reference: "Many Labs" Replication Project, *Social Psychology, 2014*.

- Replication of 13 classic and contemporary effects in psychology.
- 36 study sites (lab setting vs. online, U.S. vs. abroad) participated in this project and collected data from a total of 6,344 participants.
- Hierarchical model postulates a common τ_i among the $J+1$ studies that are labeled i.
- Keep all study sites and the effects that we can construct estimates and S.E.: 36*11 = 396 estimates.
- We consider $J = 2$ and $J = 6$. In turn, N is either 132 or 66.
- In addition to full sample, we also consider a reduced sample:

$$\begin{aligned} \text{Full samples} &: \quad J = 2,\ N = 132 \quad \text{and} \quad J = 5,\ N = 66 \\ \text{Reduced samples} &: \quad J = 2,\ N = 96 \quad \text{and} \quad J = 5,\ N = 48 \end{aligned}$$

What Are The Effects / Experiments?

2. *Gain versus loss framing (Tversky & Kahneman, 1981).* The original research showed that changing the focus from losses to gains decreases participants' willingness to take risks – that is, gamble to get a better outcome rather than take a guaranteed result. Participants imagined that the US was preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Participants were then asked to select a course of action to combat the disease from logically identical sets of alternatives framed in terms of gains as follows: Program A will save 200 people (400 people will die), or Program B which has a 1/3 probability that 600 people will be saved (nobody will die) and 2/3 probability that no people will be saved (600 people will die). In the ''gain'' framing condition, participants are more likely to adopt Program A, while this effect reverses in the loss framing condition. The replication replaced the phrase ''the United States'' with the country of data collection, and the word ''Asian'' was omitted from ''an unusual Asian disease.''

6. *Norm of reciprocity (Hyman & Sheatsley, 1950).* When confronted with a decision about allowing or denying the same behavior to an ingroup and outgroup, people may feel an obligation to reciprocity, or consistency in their evaluation of the behaviors (Hyman & Sheatsley, 1950). In the original study, American participants answered two questions: whether communist countries should allow American reporters in and allow them to report the news back to American papers and whether America should allow communist reporters into the United States and allow them to report back to their papers. Participants reported more support for allowing communist reporters into America when that question was asked after the question about allowing American reporters into the communist countries. In the replication, we changed the question slightly to ensure the ''other country'' was a suitable, modern target (North Korea). For international replication, the target country was determined by the researcher heading that replication to ensure suitability (see supplementary materials).

Estimates and Standard Errors

- In each experimental design:
  - control group of N_c participants;
  - treatment group of N_t participants.
- Each site reports the mean and standard deviation of the outcome variable for participants in each group, denoted as μ_c, μ_t, s_c, and s_t, respectively.
- **Compute the estimate** $\hat{\theta} = \mu_t - \mu_c$.
- **Compute S.E.:** $\sigma = \sqrt{s_c^2/N_c + s_t^2/N_t}$.

How Do The Estimates Across Experiments Look Like?

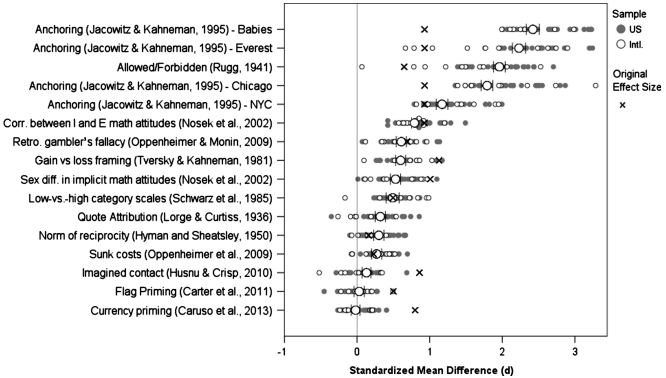
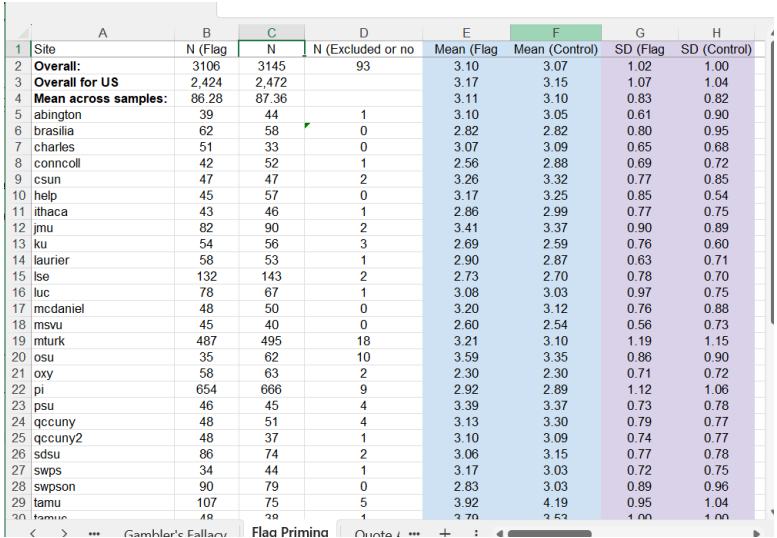


Figure 1. Replication results organized by effect. "X" indicates the effect size obtained in the original study. Large circles represent the aggregate effect size obtained across all participants. Error bars represent 99% noncentral confidence intervals around the effects. Small circles represent the effect sizes obtained within each site (black and white circles for US and international replications, respectively).

The Data We Used



	A	B	C	D	E	F	G	H
1	Site	N (Flag	N	N (Excluded or no	Mean (Flag	Mean (Control)	SD (Flag	SD (Control)
2	**Overall:**	3106	3145	93	3.10	3.07	1.02	1.00
3	**Overall for US**	2,424	2,472		3.17	3.15	1.07	1.04
4	**Mean across samples:**	86.28	87.36		3.11	3.10	0.83	0.82
5	abington	39	44	1	3.10	3.05	0.61	0.90
6	brasilia	62	58	0	2.82	2.82	0.80	0.95
7	charles	51	33	0	3.07	3.09	0.65	0.68
8	conncoll	42	52	1	2.56	2.88	0.69	0.72
9	csun	47	47	2	3.26	3.32	0.77	0.85
10	help	45	57	0	3.17	3.25	0.85	0.54
11	ithaca	43	46	1	2.86	2.99	0.77	0.75
12	jmu	82	90	2	3.41	3.37	0.90	0.89
13	ku	54	56	3	2.69	2.59	0.76	0.60
14	laurier	58	53	1	2.90	2.87	0.63	0.71
15	lse	132	143	2	2.73	2.70	0.78	0.70
16	luc	78	67	1	3.08	3.03	0.97	0.75
17	mcdaniel	48	50	0	3.20	3.12	0.76	0.88
18	msvu	45	40	0	2.60	2.54	0.56	0.73
19	mturk	487	495	18	3.21	3.10	1.19	1.15
20	osu	35	62	10	3.59	3.35	0.86	0.90
21	oxy	58	63	2	2.30	2.30	0.71	0.72
22	pi	654	666	9	2.92	2.89	1.12	1.06
23	psu	46	45	4	3.39	3.37	0.73	0.78
24	qccuny	48	51	4	3.13	3.30	0.79	0.77
25	qccuny2	48	37	1	3.10	3.09	0.74	0.77
26	sdsu	86	74	2	3.06	3.15	0.77	0.78
27	swps	34	44	1	3.17	3.03	0.72	0.75
28	swpson	90	79	0	2.83	3.03	0.89	0.96
29	tamu	107	75	5	3.92	4.19	0.95	1.04
30	tamuc	48	38	1	3.79	3.53	1.00	1.00

Gambler's Fallacy | Flag Priming | Quote /

ManyLabs Grouping

Panel A: $J = 2$

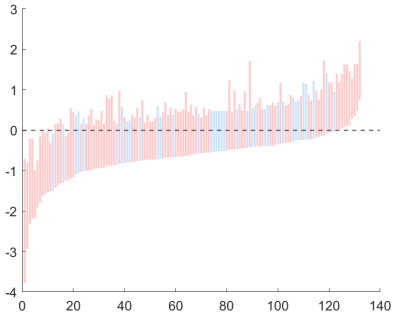
Group	Study Sites			Group	Study Sites		
1	Brasilia	Charles	Help	7	WL	TAMU	Abington
2	Laurier	MSVU	SWPS	8	QCCUNY	OSU	Luc
3	KU	SWPSON	UNIPD	9	UVA	TAMUC	PSU
4	LSE	Tilburg	WPI	10	QCCUNY2	Wisc	SDSU
5	CSUN	JMU	McDaniel	11	VCU	UFL	WKU
6	MTURK	PI	TAMUON	12	Ithaca	Conncoll	Oxy

Panel B: $J = 5$

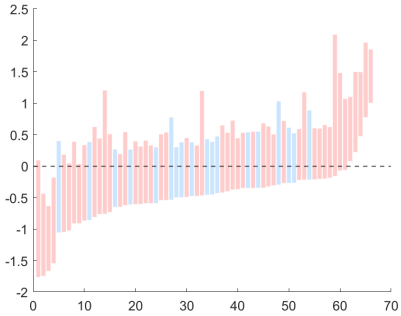
Group	Study Sites					
1	Brasilia	Charles	Help	Laurier	MSVU	SWPS
2	KU	SWPSON	UNIPD	LSE	Tilburg	WPI
3	CSUN	JMU	McDaniel	MTURK	PI	TAMUON
4	WL	TAMU	Abington	QCCUNY	OSU	Luc
5	UVA	TAMUC	PSU	QCCUNY2	Wisc	SDSU
6	VCU	UFL	WKU	Ithaca	Conncoll	Oxy

Predictive Intervals ($\hat{\nu}$, Centered and Sorted by Lower Bound)

$J = 5,\ N = 66$



$J = 2,\ N = 132$

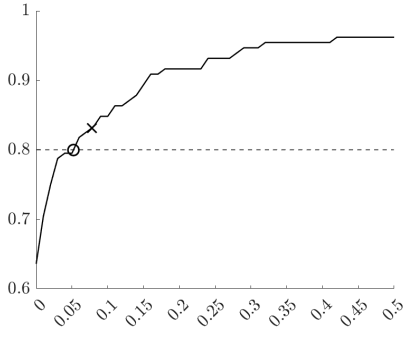


Empirical Coverage Probabilities

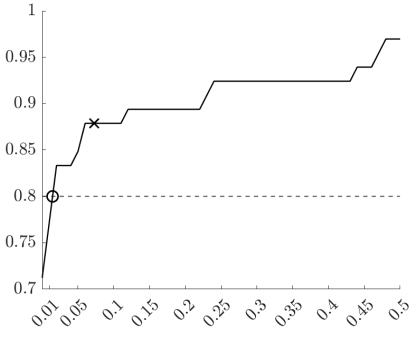
	J=2		J=5	
	N = 132	N = 96	N = 66	N = 48
	Fixed $\nu = 0$			
Emp. Coverage Freq.	0.64	0.54	0.71	0.63
Coverage p-value	0.00	0.00	0.09	0.00
	Estimated ν			
Emp. Coverage Freq.	0.83	0.82	0.88	0.81
Coverage p-value	0.51	0.60	0.13	0.87
EB Estimate $\hat{\nu}$	0.08	0.12	0.07	0.11

Many Labs Application: Coverage as a Function of ν

$J = 5,\ N = 66$



$J = 2,\ N = 132$



Somewhat Related

- Meta analysis: synthesize results from different studies. E.g. Meager (2019) "Understanding the Average Impact of Microcredit Expansions (...)," *American Economic Journal: Applied Economics*.

- External validity: are results obtained for one sample / population also valid for other population / samples? E.g., Adjaho and Christensen (2022) "Externally Valid Tretment Choice," *arXiv Working Paper*, 2205.05561v1.

- Publication bias: journals might only publish studies that generate "significant" estimates; tests and CIs loose there frequentist interpretation. E.g., Andrews and Kasy (2019) "Identification of and Correction for Publication Bias," *American Economic Review*.

- Prequential analysis: a statistical theory based on the notion of sequential predictions. E.g. Dawid (1984) "Present Position and Potential Developments; Some Personal Views: Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society A*.

- Conformal analysis: part of it is about adjusting probability statements so that they conform with empirical frequencies. Yang, Candes, and Lei (2024) "Bellman Conformal Inference: Calibrating Prediction Intervals for Time Series," *arXiv Working Paper*, 2402.05203v2.

Uncertainty in Statistical Inference

- Much of what you have seen in econometrics classes is about attaching standard errors (s.e.) to estimates.
- They can be used to construct coverage intervals, e.g.,

$$\text{point estimate} \pm 2 \times s.e. \tag{18}$$

- The s.e. are supposed to summarize uncertainty associated with your estimates.
- **We asked two questions in these lectures:** Do they? In what sense?
- Along the way, we covered:
  - Bayesian vs. frequentist inference
  - Interval forecast evaluation
  - Bayesian hierarchical modeling.